DESIGN AND PREPARATION OF THE 1996 HUB-4 BROADCAST NEWS BENCHMARK TEST CORPORA

John S. Garofolo, Jonathan G. Fiscus, William M. Fisher

National Institute of Standards and Technology (NIST)
Information Technology Laboratory
Building 225, Room A-216
Gaithersburg, MD 20899

ABSTRACT

This paper describes the procedures used in the preparation of the 1996 DARPA CSR Hub-4 Broadcast News Benchmark Test corpora and some analyses of that data. A new annotation/transcription process was designed and implemented to ensure that the transcripts[1] were practically error-free and to negate the need to hold a post-test "adjudication" as in years past. This paper focuses on this new annotation paradigm as well as an analysis of the properties of the test material.

1. BACKGROUND

The November 1995 DARPA CSR Hub-4 test was the first DARPA CSR evaluation to work with "found" speech -- speech recorded off the air and not specially contrived and collected for the test. The test data consisted of excerpts of several Public Radio International "Marketplace" radio show broadcasts.

After tabulation of the 1995 test results, interest grew in performing detailed analysis of the error rates for different data conditions. It was noted that the error rate increased substantially during areas of the test data which contained noise, music, and background speech and in areas which seemed to have been transmitted over the telephone. But the data was not annotated prior to the test in a way that would support detailed error analysis for different conditions.

The 1995 Hub-4 test data required the recognition systems to include segmentation modules since, unlike previous CSR test data, a single test set recording could contain many speakers under varying conditions. After the 1995 Hub-4 tests, there was concern that too much effort was being expended in developing condition-dependent segmentation systems and that resources should be focused on the core speech recognition technology. Further, it was thought that errors introduced by imperfect automatic segmentation could confound the interpretation of the results. And some sites did not have the resources to develop a segmenter of their own.

The analysis and segmentation challenges were dealt with by developing a segmented annotation convention and test protocol which included both a Partitioned Evaluation (PE) and Unpartitioned Evaluation (UE). In the Partitioned Evaluation, systems would be provided with segmentation information in the form of a condition tag and time stamp for each condition change. In the Unpartitioned Evaluation, systems would be required to perform their own segmentation. The data for the two tests would largely overlap so that comparisons between the PE and UE could be made.

These requirements laid the groundwork for the 1996 Hub-4 Broadcast News test paradigm and annotation convention. An enhanced annotation convention based on the 1995 Hub-4 transcript specification was first suggested. It was then decided that a more portable and extensible SGML-based convention should be adopted. An SGML convention was designed by the LDC [Graff] and then iteratively refined via input from NIST and George Doddington. The annotation convention was constructed to enable a number of conditions to be identified and time-segmented. Particular combinations of these conditions were chosen as "focus conditions" and were used to create the PE test index and in partitioning the PE and UE evaluation test scores. [Pallett] The following table shows the 1996 Hub-4 focus conditions and associated condition combinations:

Condition	Dialect	Mode	Fidelity	Background
Baseline Broadcast (F0)	native	Planned	High	Clean
Spontaneous Speech (F1)	native	Spontaneous	High	Clean
Reduced Bandwidth (F2)	native	(any Mode)	Med/Low	Clean
Background Music (F3)	native	(any Mode)	High	Music
Degraded Acoustics (F4)	native	(any Mode)	High	Speech/Other Noise
Nonnative Speakers (F5)	nonnative	Planned	High	Clean
All Other Combinations (FX)	–	–	–	–

Table 1: 1996 Hub-4 Focus Condition Definitions

The 1996 Hub-4 training and development test data was collected and annotated at the LDC using the new convention and special tools. [Graff]

Prior to the collection and annotation of the evaluation test data, it was observed that the training data contained a number of annotation/transcription errors - many due, no doubt, to the complexity of the new annotation task. We were concerned that if a similar error rate occurred in the evaluation test transcripts, that it could jeopardize the accuracy of the tests and result in a lengthy adjudication process. George Doddington suggested that if the transcripts were separately annotated and transcribed by 2 or more people and reconciled, that "perfect" transcripts could be produced. Such a multi-pronged annotation approach was too

[1] "Transcript" here refers to the document containing both annotation and orthography.

expensive to be applied to the training data, but was deemed to be worthwhile for the evaluation test data - especially if a formal adjudication process could be avoided.

Such a process would provide not only much cleaner transcripts, but it would also allow us to perform a small-scale experiment in transcriber error and agreement.

2.0 BENCHMARK TEST DATA PREPARATION

We formulated a data preparation plan which involved producing the final evaluation transcripts using a pipelined 4-stage process. Rather than having a single annotator or team transcribe the entire test set as had been done in the past, three annotators would produce parallel transcripts. In order to avoid the extremely difficult task of reconciling three separate sets of transcripts complete with annotations and orthography, the transcripts would be generated in stages with reconciliation between the parallel versions occurring between each stage.

First, the three annotators would each generate the SGML-tagged annotations for the test set. The three versions of the annotations would then be reconciled by a fourth person. The reconciled annotations would then provide the framework for transcription of the orthography by the three annotators. Finally, the three versions of the orthography would be reconciled to produce a clean annotated transcript.

This staged approach facilitated transcript reconciliation and permitted us to release the segmentations to the test participants for the PE test before the final transcriptions were completed. It also allowed us to determine if material was lacking for any particular PE test condition and add material if necessary.

2.1 Selection

The LDC provided an initial segmentation of a pool of about ten hours of potential test material. NIST used the LDC annotations in conjunction with its own ad hoc analysis of the data to select the UE test set. Initially, a set of four shows was chosen to evenly represent television and radio broadcasts. Two of these shows ("PRI Marketplace" and CSPAN "Washington Journal") were represented in the training and/or development test set and the remaining two shows (NPR "The World" and CNN "Morning News") had not been previously used in system development. Contiguous half-hour excerpts were then chosen from these shows on the basis of probable focus condition coverage.

2.2 Annotation

The annotation process was carried out by one representative from NIST and one representative from NSA and a group of annotators at the LDC. The annotators used the LDC annotation tool, "Hubsaver", [Graff] and were each given a dedicated workstation for the task. The annotators were given the standard LDC Broadcast News annotation instructions and the Hub-4 annotation document and instructed to not discuss the annotations with each other, although they were permitted to discuss and ask questions regarding general annotation rules.

The process of annotating the four half-hour broadcast segments took approximately five working days for each annotator to complete - approximately twenty times real time.

2.3 Annotation Reconciliation

Once the annotations were complete, they were adjudicated by a reconciliation tsar. The reconciliation tsar used a NIST/LDC - created tool which took as inputs time-aligned SGML tags and output a single tag set. Figure 1 shows an example of the tool being used to reconcile a Segment tag.

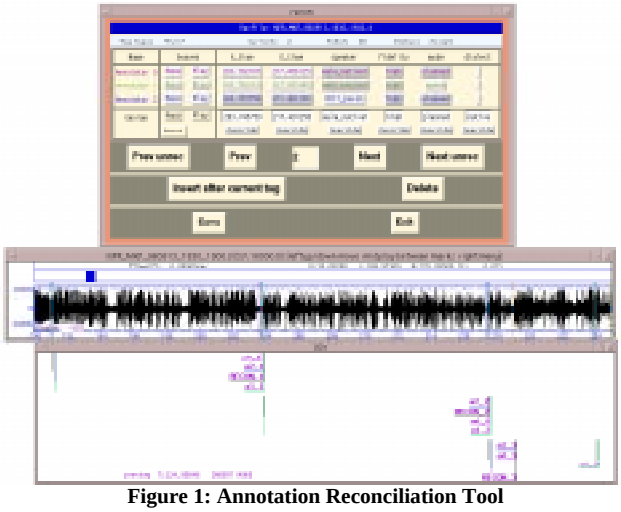

Figure 1: Annotation Reconciliation Tool

To reduce complexity, one tag type (e.g., Session, Segment, Background, etc.) was processed at a time. In the course of reconciling a single broadcast, many passes were made. After all the tag sets were completed for a broadcast, they were re-integrated using PERL scripts. The tag sets provided the temporal framework for transcription of the orthography.

Annotation reconciliation was more time-consuming than that of annotation and took approximately 8 working person-days to complete.

2.4 Transcription

Each of the annotators were given the reconciled annotations as a basis for transcription. They were instructed to not change any tags and if they found what they believed to be a tag error, they were to make a note of it. By the time transcription began, the tags had already been released for the PE test and were "locked".

During the transcription process, 30 annotation tags were found to contain errors. Of these, 2 resulted in a change in Partitioning in the test data. The effected Partitions, amounting to 1.2 seconds of data, were tagged for exclusion from scoring.

The transcribers again used the LDC Hubsaver tool for transcription. The transcription process took approximately 9 days per transcriber to complete. Note that the LDC transcribers, who were more experienced with Hubsaver and the Broadcast

News annotation/transcription convention, may have been a bit faster than the NSA and NIST transcribers.

2.5 Transcription Reconciliation

Once the transcriptions were complete, each segment was reconciled by one of two transcription tsars. The three sets of transcriptions were merged and differences highlighted as alternations. The transcription tsars used the merged transcriptions and LDC Hubsaver tool to reconcile the differences.

A single transcription was always chosen by a tsar; no alternation convention was used. Each ambiguous case was decided in favor of what the tsar believed the speaker meant to say, rather than what was actually said.

The process of reconciling the 1622 transcription differences took approximately 4 person-days to complete, giving an average rate of a little under one difference reconciliation per minute.

2.6 Orthographic Transformations

As in previous years' evaluations, we applied a set of global mapping rules to both REF and HYP transcriptions in order to wipe out some differences that we thought should not be counted as errors. The rules fall into four classes: contractions, alternate standard spellings, spelling errors in training transcriptions, and compound words.

2.6.1 Contractions

Contractions were handled by marking each occurrence of a contraction in the REF files with a special marking that functioned as an alternation of the contraction or the words it contracted *in that context*, then pre-processing the HYP files with a set of mapping rules that replaced each contraction in them with an alternation of all the strings of words that it might contract, e.g.

JOHN'D => {JOHN HAD / JOHN WOULD}

Thus when the scoring was done,

- if a contractable string of words was said, either that string of words or one of its allowable contractions would be counted as correct;
- if a contraction was said, either the contraction or the single string of words that it contracted in that particular utterance would be counted as correct.

In order to make up the mapping rules, a list was first made of all words in the 1996 Hub-4 Training, Development Test, Evaluation Test, and recognition system output transcripts that included an apostrophe and then examined by hand to throw out non-contractions such as "O'Reilly". This yielded a list of 901 contractions. These contractions and other linguistic sources were then examined in order to find generalizations that were incorporated into a program that produced a mapping rule to expand each one. These mapping rules were then examined and edited by hand.

It is important to note that if the contraction-expansion rules overgenerate, that is, produce an uncontracted form that the contraction in its context could not have, the evaluation system will not detect a real error in some cases.

Consider "X's". In general, "X's" can represent either the inflected possessive form of X, "X is", or "X has". That simple a rule, though, will greatly overgenerate, since in most particular contexts "X's" cannot be a contraction for one or more of these possibilities. It cannot represent "X has", for instance, if "has" is the main verb, as in "John has a dog", but only if "has" is an auxiliary verb, as in "John has gone home"; it cannot represent the possessive in "John's gone home".

If our mapping rules allow the expansion
"JOHN'S => JOHN HAS",

then every REF case of the form
NP has NP (e.g. "John has a dog")

will mistakenly match with no error the HYP transcription
NP's NP (e.g. "John's a dog").

If we *don't* allow "JOHN'S => JOHN HAS", then

every REF case of the form
NP HAS PAST_PARTICIPLE (e.g. John has gone crazy.")

will mistakenly count an error in the HYP:
NP'S PAST_PARTICIPLE (e.g. John's gone crazy.")

The task of writing simple literal rewriting rules to expand apparent contractions with minimal over-generation turned out to be very difficult, and the true rules are a good deal more complicated than the ones we had time to develop. The basic problem is that a correct rule to expand contractions must be sensitive to at least syntactic structure, in ways that are sometimes subtle and not well known. We believe that an algorithm sensitive to part-of-speech tags could be made that would expand contractions with little or no overgeneration, and we have already collected some of these generalizations.

We would like for the rules to not apply to any possessive forms, but in fact our rules over-generate by applying to some possessives like "John's hat" as well as all contractions. In order for the possessives not to be counted as errors, the rules for "X's" also output the original one-token possessive form.

2.6.2 Alternate Standard Spellings

Another subset of 63 mapping rules allowed alternate standard spellings. While our major source for alternate spellings was the American Heritage Dictionary, we did use Web searches in quite a few cases, particularly to find alternate spellings of people's names.

In some cases the rules were context-sensitive; for instance, "Falkner" and "Faulkner" are valid alternate spellings only if the American writer William Faulkner is being referred to.

2.6.3 Transcription Errors In The Acoustic Training Data

Since the acoustic training data that was provided included some errors in transcription, we decided to forgive certain substitution errors if the hypothesized token occurred in the training data as one representation of the word that was actually said. In order to do that, we made a concentrated effort to find all errors of transcription in the training data (here including the development test).

Several semi-automated tools were used to find transcription errors. First off, in a straightforward spell-checker approach, all tokens which were OOV relative to a large standard lexicon were put onto an exception list ("hot list"), whose items were then examined to see if they were in fact errors. Our second approach was similar, except that the hot list consisted of word occurrences that were merely highly improbable, sorted by probability, using both unigram and trigram probabilities relative to a statistical language model. A third avenue used our "phonological distance" function to compute near homophones: a word occurrence was added to the hot list if it was quite probable in a context but had a similar-sounding word that would have been very improbable in that context. (This last approach explicitly models errors as being produced by transcribers who are over-biased toward the word they expect to hear, rather that the word that is actually said.) Another pass at the problem consisted of examining in context all occurrences of certain words that are notoriously confusable, such as "counsel/council" and "affect/effect". And a last approach simply applied a set of rules that correct common misspellings when the error is obvious, as in "commitee", "hte", "Phillipines", etc.

Each transcription error thus found produced a case of the same word being represented two different ways in the official training data, and a rule was made up to map one of the spellings into the other, for possible use in forgiving errors during evaluation. In all, there were 411 of these rules (note that the number of error occurrences would be considerably higher than this). The decision was made to use only the rules covering mistranscriptions that did not result in a standard spelling of a different word, with the exception of compound-word rules, reducing the size of the final set of rules used to 348.

2.6.4 Compound Words

Compound words that have a unique standard spelling (most of them, under our assumption of one dictionary as the primary standard) were handled along with other words as possible cases of mistranscription in the training data, except that the exemption for misspellings that resulted in valid words did not apply to them. In other words, the rules forgave misspellings of them only in case the misspelled form had occurred in the training data.

There was a small set of eleven compound words for which a standard spelling was uncertain, such as "WEBSITE/WEB SITE", and mapping rules were made up to allow those variations.

3.0 PROPERTIES OF THE EVALUATION TEST SET

3.1 Focus Conditions

This year, additional material was not added to the UE test set to create the PE test set, so the UE and PE test sets are identical in content. With the exception of Focus Condition 5 (non-native speakers), the coverage for all focus conditions for the Evaluation Test set was good [Figure 2]. Note that because the combination of conditions required for F5 is rare, the amount of F5 material in the test pool (and in the selected test set) was limited.

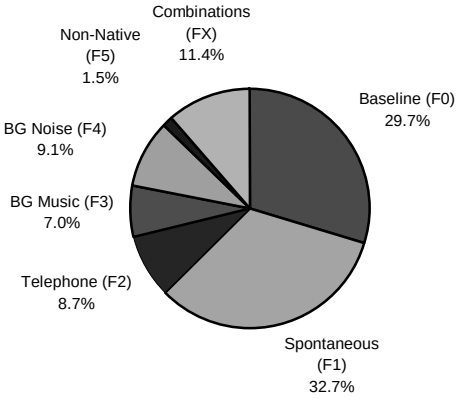

Figure 2: Hub-4 1996 Evaluation Test Data Distribution by Focus Condition (in words)

In Figure 3, we compared relative percentages of the distribution of focus conditions between the Evaluation Test Set, Development Test Set, and the training data. We found that the relative distribution of Focus Condition F5 in the Evaluation Test Set is very similar to that of the training data, but not the Development Test Set. This is because the Development Test Set was supplemented with additional F3 and F5 data for the PE portion of the test. We suspect that the remaining differences are due to the small sample size in the test sets.

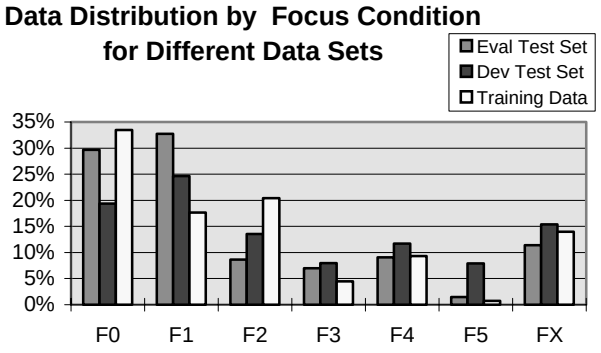

Figure 3: Evaluation Test, Development Test, and Training Data Distribution by Focus Condition (in words)

3.2 Some Other Characteristics

We analyzed these segment-averaged characteristics of the development test and evaluation test transcriptions to see if there were significant differences in distribution:

- ROS2 - rate of speech in syllables/sec., computed as the number of REF syllables divided by the duration of speech.
- FRAG1 - rate of fragmentation in number of fragments per 1000 words.
- OOV1 - Out-of-vocabulary rate per 1000 words relative to a CMU language model.
- PERP1 - test set perplexity relative to the CMU language model.

In all cases, the Mann-Whitney U test was used to assess significance. The only one of these four characteristics showing a significant difference between Development and Evaluation test sets was ROS2, shown in Table 2 below:

Data Sets	N1/N2	Mean	Z	P(Z)
TRN/DEV	10343/561	4.5/4.7	-3.05	.0022
DEV/EVL	561/381	4.7/4.9	-3.49	.0004
TRN/EVL	10343/381	4.5/4.9	-7.04	.0001

Table 2: ROS2 Values and Differences.

The increase in ROS2 from training set to Development test set to Evaluation test set may have been caused by a decrease in transcription error. We did more stringent verification and correction of the Evaluation test set than of the Development test set, and very little or none on the training data. The differences in ROS2 between the Development test set and the Evaluation test set are concentrated in Focus Condition F3 and, to a lesser degree, F4. These are difficult listening conditions, and under such conditions, transcriber errors that change the word count probably tend to be deletions rather than insertions. Thus the more correct transcriptions would include more words and have a higher indicated rate of speech.

4.0 ANALYSIS OF ANNOTATION/ TRANSCRIPTION RECONCILIATION

After the final transcripts were complete, we were able to use data collected in annotation, transcription, and reconciliation to examine annotator/transcriber error rates and agreement. Although many human recognition performance experiments have been performed in the past, it is interesting to look at annotator/transcriber performance in the context of the Hub-4 tests. Unlike in many other human recognition experiments, the transcribers were not under time pressure and could back up and re-play any portion of the recording and review their annotations and transcriptions at any time. They were also at liberty to use dictionaries and other "performance-enhancing" tools. So, theoretically, they should have achieved the best possible performance. This experiment quantifies the difficulty in generating a canonical Hub-4 transcript and highlights the subjectivity involved in condition labeling and segmentation.

4.1 Analysis of Annotation Reconciliation

While many different conditions were annotated in the transcripts (Section, Speaker, Mode, Non-native, Fidelity, etc.), we chose to focus our analysis on the background noise annotations which we believe to be the most problematic to annotate. And correct classification of background noise is crucial since background conditions differentiate Focus Conditions F3 and F4 from Focus Conditions F0, F1, F2 and F5.

Three background noise conditions are annotated: music, speech, and noise (sometimes referred to as "other"). Background noise is a catch-all classification that identifies noise other than speech or music. Speech from a single talker is considered background speech while speech from more the one talker is considered background noise. Table 3. shows the total time for each of the three background conditions in the evaluation test set.

	Total Test Set	BG_Music	BG_Speaker	BG_Noise
Time (Sec):	6544.40	693.16	218.07	466.54

Table 3: Test Set Background Noise Distribution in Seconds

By treating background condition annotation as a detection task, we can compute rates of missed detection and false alarms for each annotator and each background condition. For the annotation task, we defined the *missed detection rate* to be the ratio of the duration of areas which should have been labeled for a particular background condition, but were not to the actual total duration of the background condition as determined after reconciliation. For example, the missed detection rate percentage for annotator 1 for background music is [(144.72 /693.16) * 100] or 20.9.%

The *false alarm rate* is defined to be the ratio of the duration of areas which should **not** have been labeled for a particular background condition, but were, to the actual total duration of areas without the background condition as determined after reconciliation. Again, for example, the false alarm percentage for annotator 1 for background music is [(24.74 / 5851.2) * 100] or 0.4%.

Missed detection and false alarm percentages quantify the annotator's labeling inaccuracy, but they do not express the total duration of incorrectly labeled background conditions relative to the entire test set. Thus, a third metric is required to present a complete picture of annotation accuracy. *Annotation error rate* is the ratio of the sum of areas incorrectly labeled for background condition to the total duration of the test set. For example, the annotation error percentage for annotator 1 for background music is [((144.72 + 24.74) / 6544.40) * 100] or 2.6%.

Table 4 depicts each of these measures for each annotator for each background condition. Note that the annotators did best at classifying background music, followed by background speaker and background Noise.

Annot	Background Music			Background Noise			Background Speakers		
	Missed Detect	False Alarm	Error Rate	Missed Detect	False Alarm	Error Rate	Missed Detect	False Alarm	Error Rate
1	20.9	0.4	2.6	57.0	33.7	35.3	87.9	0.0	3.0
2	8.5	10.2	10.1	56.7	3.4	7.2	25.4	0.5	1.3
3	4.5	0.4	0.9	68.0	1.8	6.6	75.7	0.0	2.6

Table 4: Percent Annotator Accuracy for Background Conditions

The table indicates that all 3 annotators had difficulty detecting background noise and background speakers. It is interesting to note that annotator 1 also had a high false alarm rate, and therefore, also a high error rate for noise. Annotator 1 also had difficulty detecting background music. The other error rates were generally low.

It would be interesting to compare these to the output provided by an automatic noise classification and segmentation system. The variable error measures may have been due to either insufficient training or an ill-defined task. We believe that the latter accounted for most annotation errors. For example, the determination of conditions such as noise levels and perceived bandwidth are highly subjective and the annotators expressed these difficulties in discussions.

4.2 Analysis of Transcription Reconciliation

We performed two analyses of transcriber performance for the transcription phase. The first involves a tabulation of agreement between each transcriber for each Focus Condition. The second analysis compares each transcriber's transcriptions to the final reconciled transcriptions after applying all of the lexical mapping rules and transformations used in the Hub-4 evaluation. This analysis allows comparison with the output of the evaluation speech recognition systems.

In order to compute transcriber agreement, a tool was written using the SCLITE scoring package which aligned any number of parallel transcriptions together into a single network of tokens. The transcribers were said to agree where each of them supplied identical tokens which aligned to each other. Since raw counts of agreed-upon words would be difficult to interpret, we produced a percentage by dividing the sum of agreed-upon words by the number of transcription tokens in the reconciled transcripts.

Figure 4, summarizes the inter-transcriber agreement on the transcribed tokens. The bars coded, "A1 vs A2 vs A3", represent the 3-way percent agreement between transcribers. The over-all 3-way agreement percentage of 89% indicates that the reconciliation tsar had to resolve 11% of the tokens. The highest agreement was found surprisingly in focus condition F5 (Non-native) which is perhaps due to its small sample size. Not surprisingly, F0 (Baseline) also had high agreement percentages.

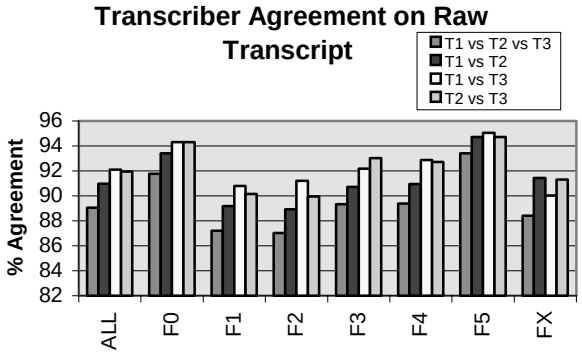

Figure 4: Transcriber Agreement

The second transcriber comparison was to compare each transcriber's transcription to the final reconciled transcripts. We treated each initial transcript as though the transcription had been generated by a recognizer, and scored them against the final reconciled reference transcript. The transcripts were scored using the same filtering and scoring procedures as applied to the official benchmark test submissions, with one exception -- reference word fragments were not alternated with NULL, so the transcriber was forced to correctly transcribe them. Figure 5 summarizes the word error rates for each of the transcriber's transcriptions by Focus Condition. We have also shown the best recognizer output (R-best) for each Focus Condition from the Hub-4 tests. [Pallett] Again, focus condition F5 stands out with the lowest word error rates for all three transcribers, followed by F0. The lowest word error rate (0.3%) was achieved by Transcriber 2 for F5. The highest word error rate (5.4%) was from Transcriber 1 for FX. More interesting, perhaps, is the large difference between recognizer performance and human transcriber performance. Broadcast news recognition systems must improve substantially if they are to approach the performance of even an "imperfect" human transcriber.

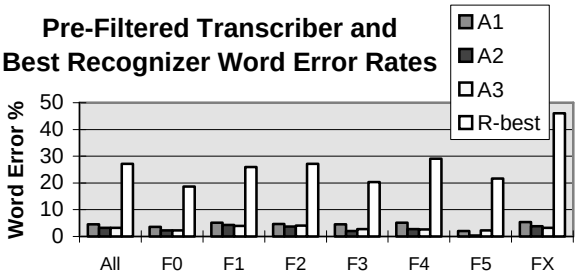

Figure 5: Transcriber/Best Recognizer Word Error Rates

The annotation phase was more problematic for the annotators than the transcription phase as indicated by the annotators' comments and significantly lower error rates for transcription than annotation. Note that while transcriber agreement regarding the orthography usually exceeded 90% and word error rates were usually less than 5%, the missed detection, false alarm, and error rates for annotation varied widely and frequently exceeded 50%, occasionally even approaching 100%. It is clear from this that the annotation aspect of the Hub-4 transcript preparation task is more difficult and/or ill-defined than the transcription task. The

annotation task needs to be re-examined, and procedures and specifications clarified or modified.

5.0 SUMMARY

The new Broadcast News annotation convention has been a valuable resource in permitting partitioned recognition and analysis of Broadcast News corpora. It has allowed us to focus on specific corpora conditions and develop improved scoring and data properties analysis. But the improvement has not come without cost. The preparation of this year's evaluation test set was time-consuming and expensive. If we had not developed and implemented the strategy we did to insure the correctness of the transcripts, the entire community would have borne the cost during a difficult post-test adjudication.

The transcript preparation strategy we employed had some obvious benefits:

1. The transcripts were "clean" prior to the test, so sites did not have to spend time writing bug reports for adjudication..
2. The correction of errors in the data was not biased toward decreasing test error rates as in the past.
3. By eliminating adjudication, we were able to push back the evaluation schedule and permit sites more development time.

We believe that this resulted in a higher quality and fairer test. But the strategy had some significant costs:

1. The test schedule was strongly front-loaded this year and we had just barely enough time to prepare the test set prior to the test start date.
2. The test data preparation was more costly than in previous tests – about 400 labor-hours by a rough calculation.

Our analysis of the multiple annotations suggests that the many inter-annotator inconsistencies may be caused by the complexity of the task and lack of definition of the categories. Our annotation reconciliation tsar sometimes had difficulty determining the correct annotation even when provided with three possible annotations. A simpler annotation convention would help, as would defining category boundaries with examples. The addition of automatic tools which provide some relevant characteristics such as bandwidth and noise might also be helpful. Finally, the use of an improved annotation and transcription tool with tighter integration of segmentation and orthography would help reduce errors and speed all phases of transcript preparation.

For next year, we are considering streamlining the annotation process by starting with a single set of annotations which would then be verified and corrected by an annotation tsar. We believe that we can still attain a level of accuracy comparable with this year's annotation process and reduce the time and difficulty in reconciliation. We believe that the transcription process used this year worked well, so we intend to use the same process next year.

ACKNOWLEGEMENTS

The authors would like to thank David Pallett for his input on this paper. We would like to thank George Doddington for his help in establishing the design of the Broadcast News annotation convention and in formulating and helping with the annotation and transcription reconciliation process. We would also like to thank Shirley Ramsey and Greg Sanders for their roles as tireless annotators and transcribers. Finally, we would like to especially thank Dave Graff and the LDC annotators for their hard work and dedication in preparing the training and development test material and for helping with the preparation of the evaluation test material and Zhibiao Wu for his help in the creation of the annotation reconciliation tool.

NOTICE

REFERENCES

Graff, D., Wu, Z.., MacIntyre, R.., Liberman, M., *The 1996 Broadcast News Speech and Language-Model Corpus*, Proc., 1997 DARPA Speech Recognition Workshop.

Pallett, D.S., and Fiscus, J.G., *1996 Preliminary Broadcast News Benchmark Tests*, Proc., 1997 DARPA Speech Recognition Workshop.